REAL-TIME DEEPFAKE DETECTION FOR ONLINE MEDIA INTEGRITY

AGNAL MENACHERY (JEC20AD003)
NIKHITHA JOY(JEC20AD036)
PRADUL O P(JEC20AD038)
SREELAKSHMI SUDHEER(JEC20AD049)

Supervised by: Mr. BINEESH M

Department of Artificial Intelligence and Data Science,
Jyothi Engineering College, Cheruthuruthy.

INTRODUCTION

- Deepfakes are a growing menace in the digital world.
- Aims to tackle the rising threat of deepfake contents by developing an advanced real-time detection system.
- Leveraging state-of-the-art algorithms, our system stays ahead of evolving deepfake techniques, providing robust defense against manipulation.

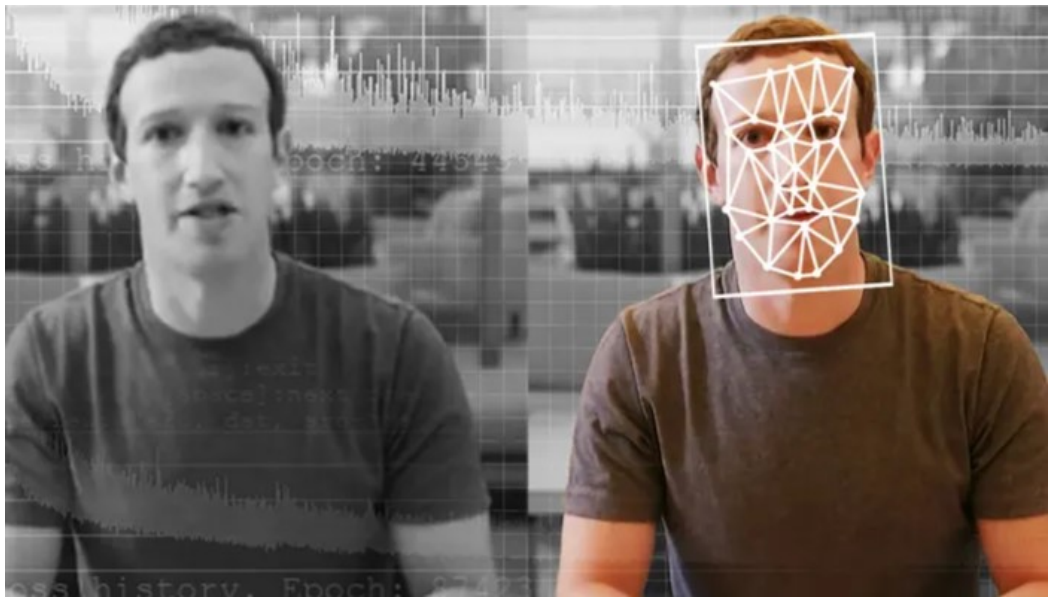

Figure 1: Deepfake Detection

BACKGROUND

Kerala man loses Rs 40,000 to AI-based Deepfake WhatsApp fraud, all about the r scam

In a recent case of an online scam on WhatsApp technology to dupe Rs 40,000 from a man from

Deepfake scam: Company lo fake video call from 'CFO'

The employees that attended this conference call were instructe different Hong Kong Bank accounts.

Viral Video Of Actress Rashmika Mandanna

Deepfake: Actress Alia Bhatt morphed video gets attention online; increases concern over misuse of

Viral Deepfake Videos Thrive Of Aamir Khan & Ranveer Singh Endorsing Political Parties

LITERATURE SURVEY

- Various techniques for deepfake detection - human-machine collaboration [1], hand-crafted features [2], and deep learning approaches [3].
- Technical deepfake detection methods like PRNU analysis [4], battleground dynamics [5], and artifact identification via deep learning[6], stressing the need for reliable authentication and ongoing innovation.
- Innovative approaches for deepfake detection, including MINTIME [7] and a Convolutional Vision Transformer [8].
- [9], [10], [11] emphasizing fusion methods, comprehensive reviews of deepfake detection techniques using deep learning
- CLRNet [12], error-level analysis[13], Mesonet[14] provide innovative deep learning-based methods for deepfake detection.
- Temporal facial features and adversarial training [15].

GAPS IDENTIFIED

1. Data limitations can hinder the AI's ability to recognize the wide range of deepfake variations.
2. Using GANs to generate highly realistic content,making it difficult for AI to distinguish between real and fake.
3. Lack generalizability and do not utilize temporal information.
4. Data quality issues.

PROBLEM STATEMENT

- To develop a robust and real-time DeepFake detection solution for enhancing the security and trustworthiness of online media in digital communication, including social media posts, video calls, and various other forms of digital content.

OBJECTIVES

- To develop an advanced deep learning algorithm to detect deepfake videos in real-time.
- To optimize resource efficiency, to develop a deepfake detection model that can effectively operate on simple systems.
- To conduct comprehensive testing to ensure the detection model accurately identifies deepfakes.
- To integrate the deepfake detection model to a user interface.

METHODOLOGY

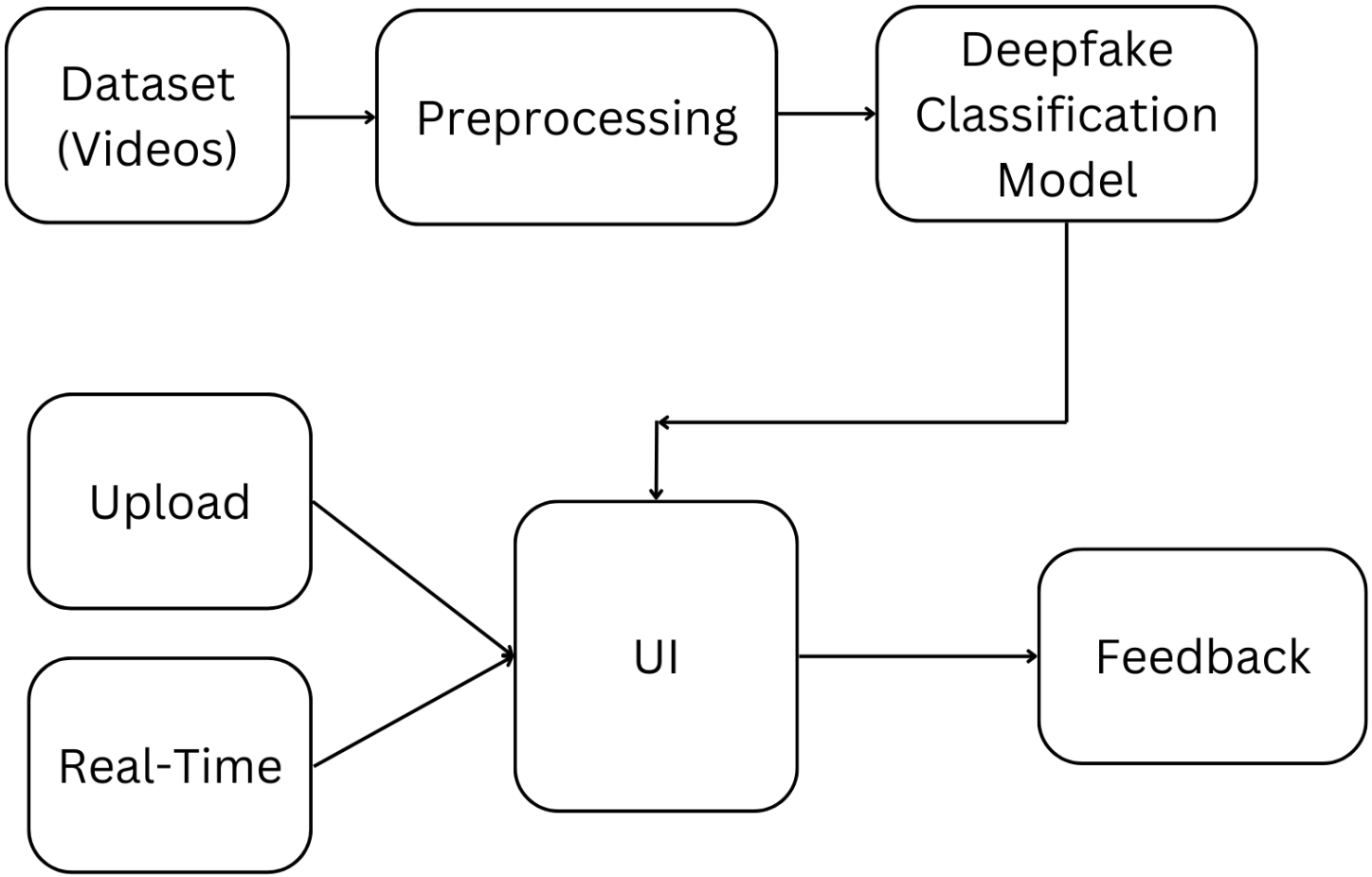

Figure 2: Methodology

BEHIND THE MODEL

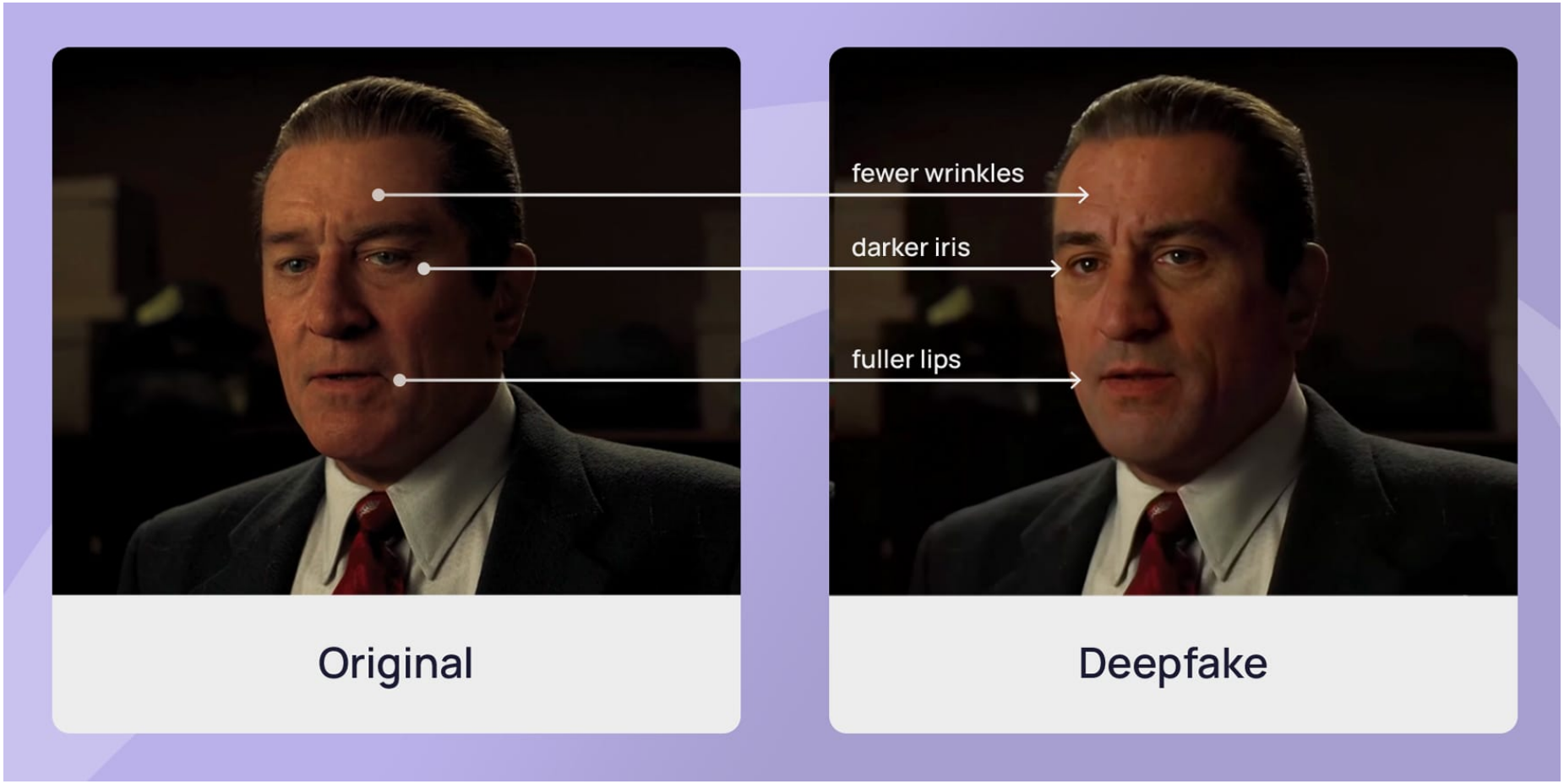

Figure 3: Difference between original and deepfake

DATASET

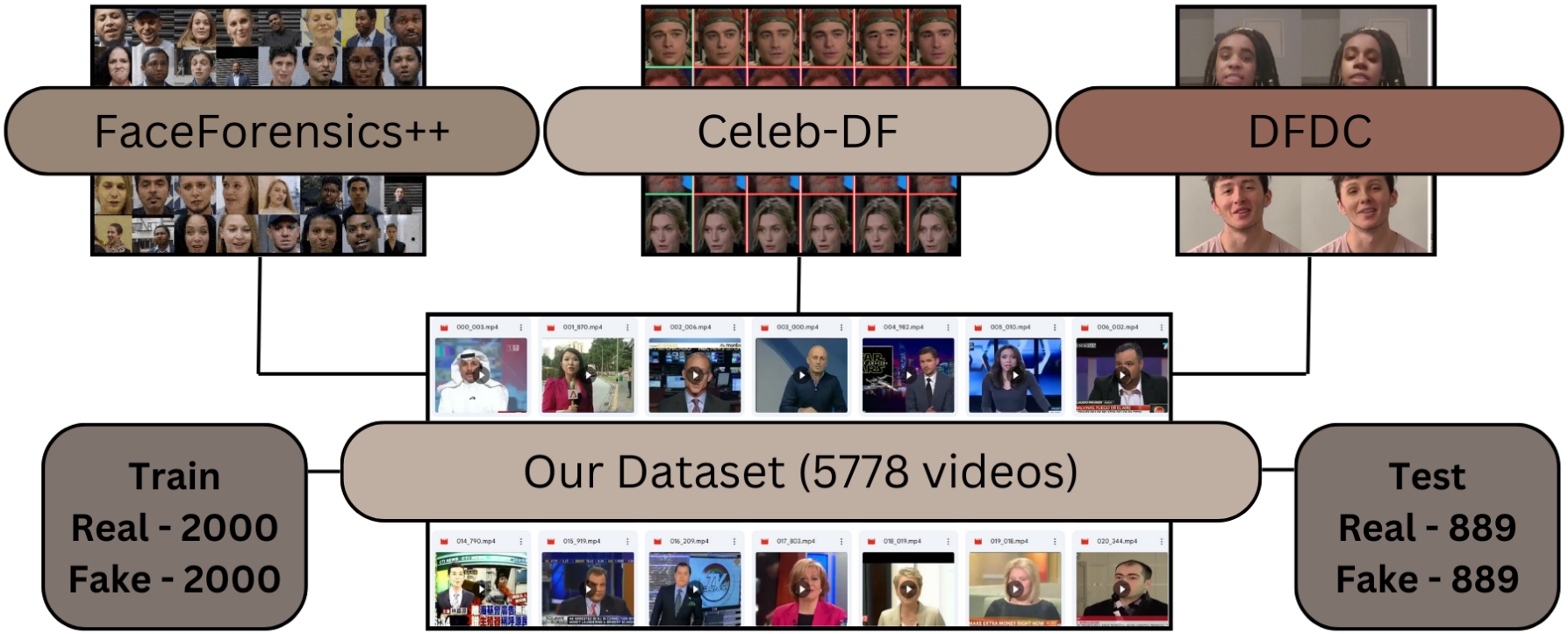

Figure 4: Dataset

MODEL ARCHITECTURE

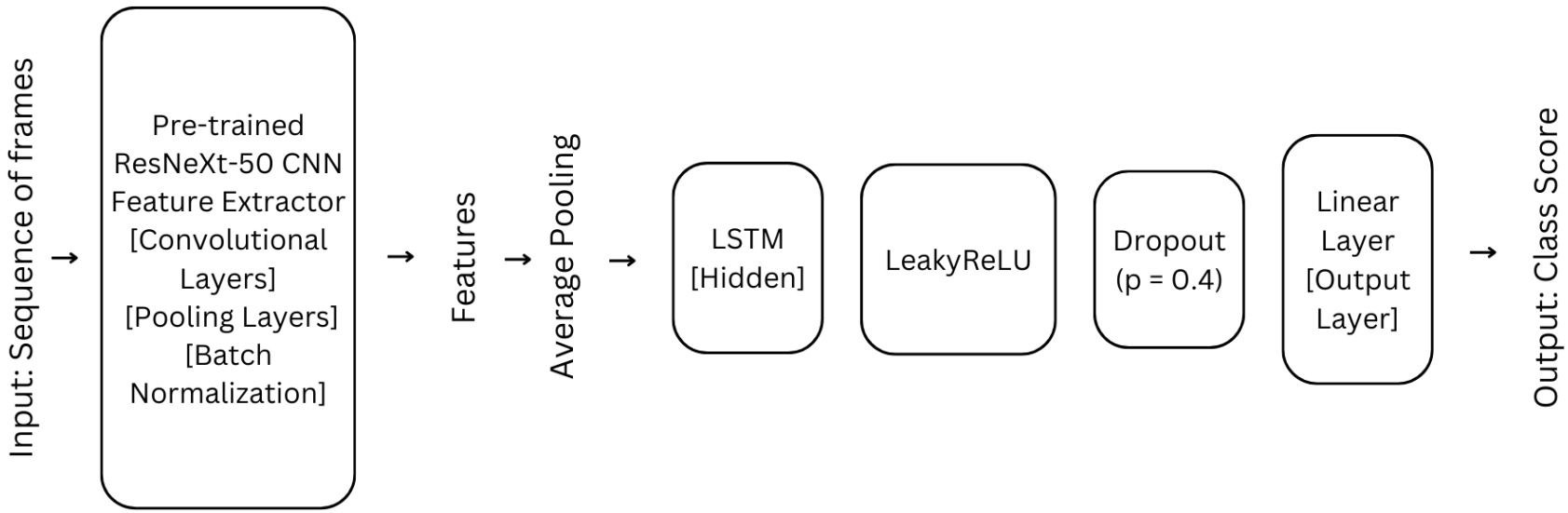

Figure 5: Model Architecture

RESNEXT ARCHITECTURE

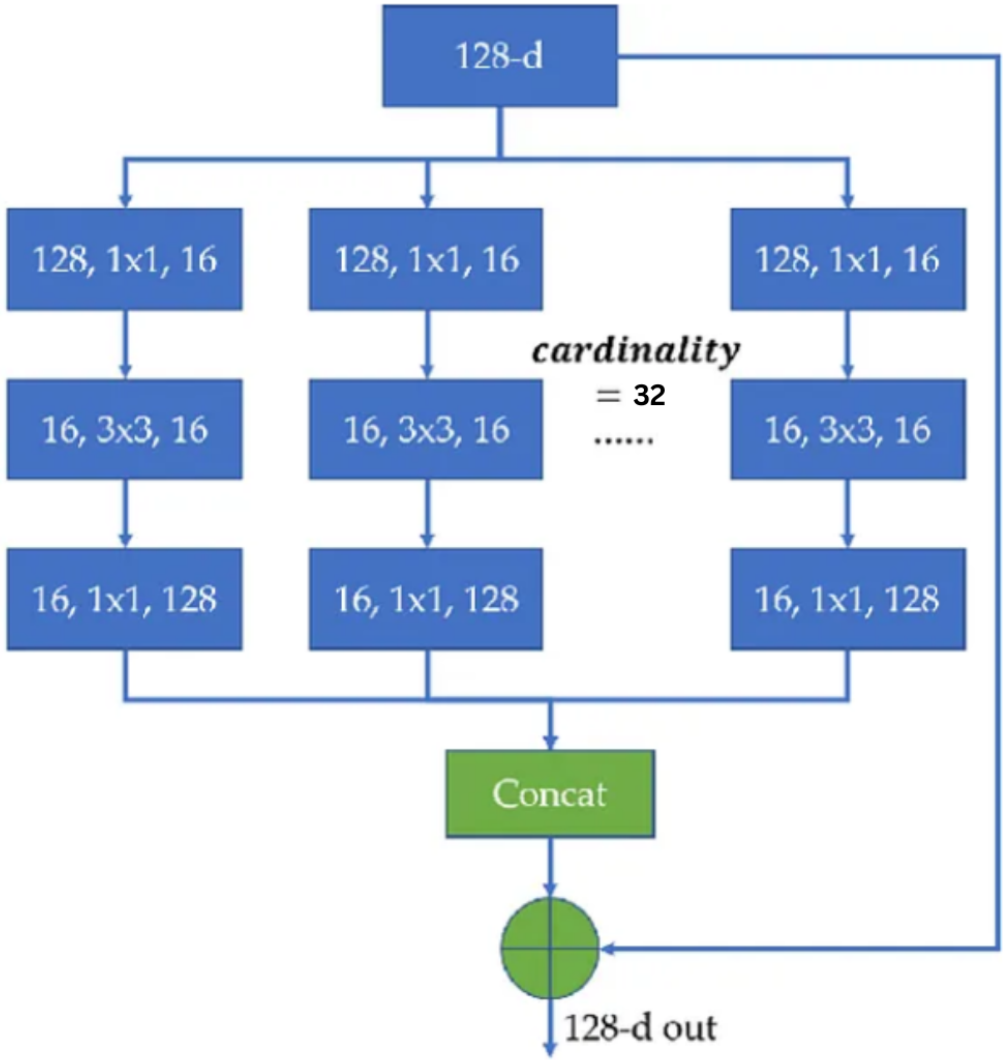

Figure 6: ResNeXt50_32x4d [16]

LSTM ARCHITECTURE

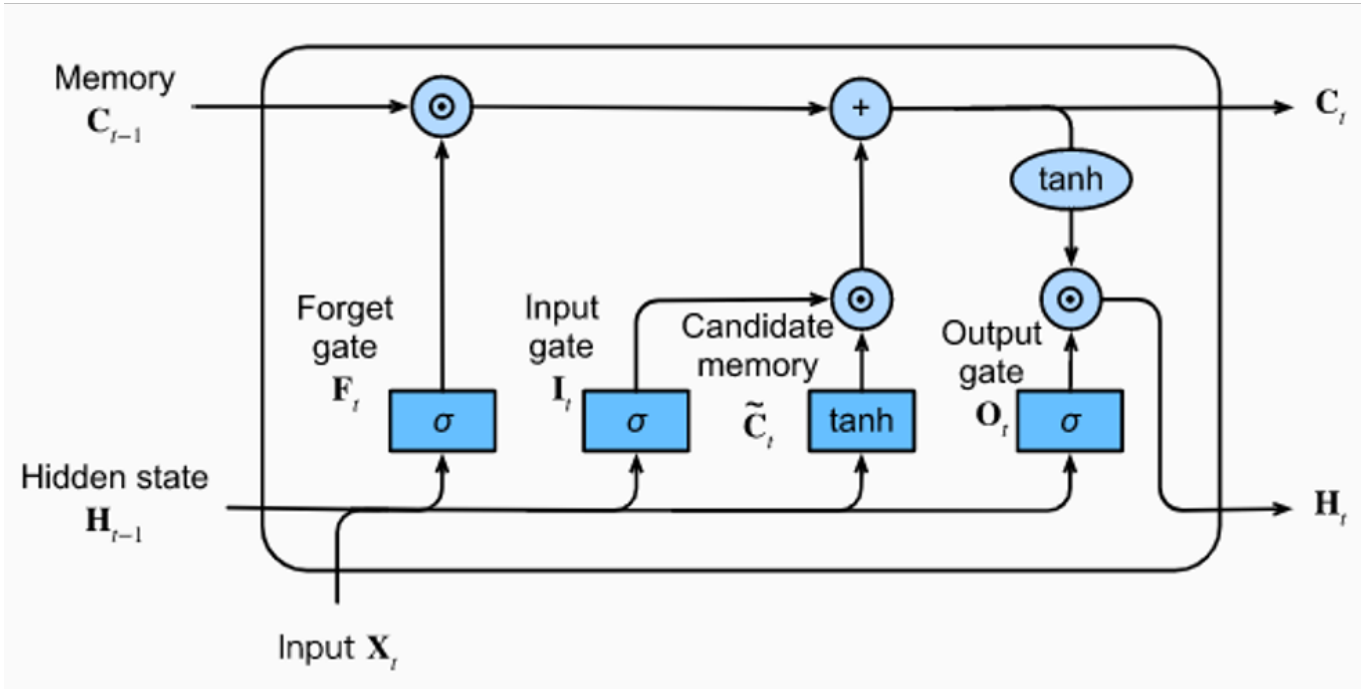

Figure 7: LSTM Cell

USER INTERFACE

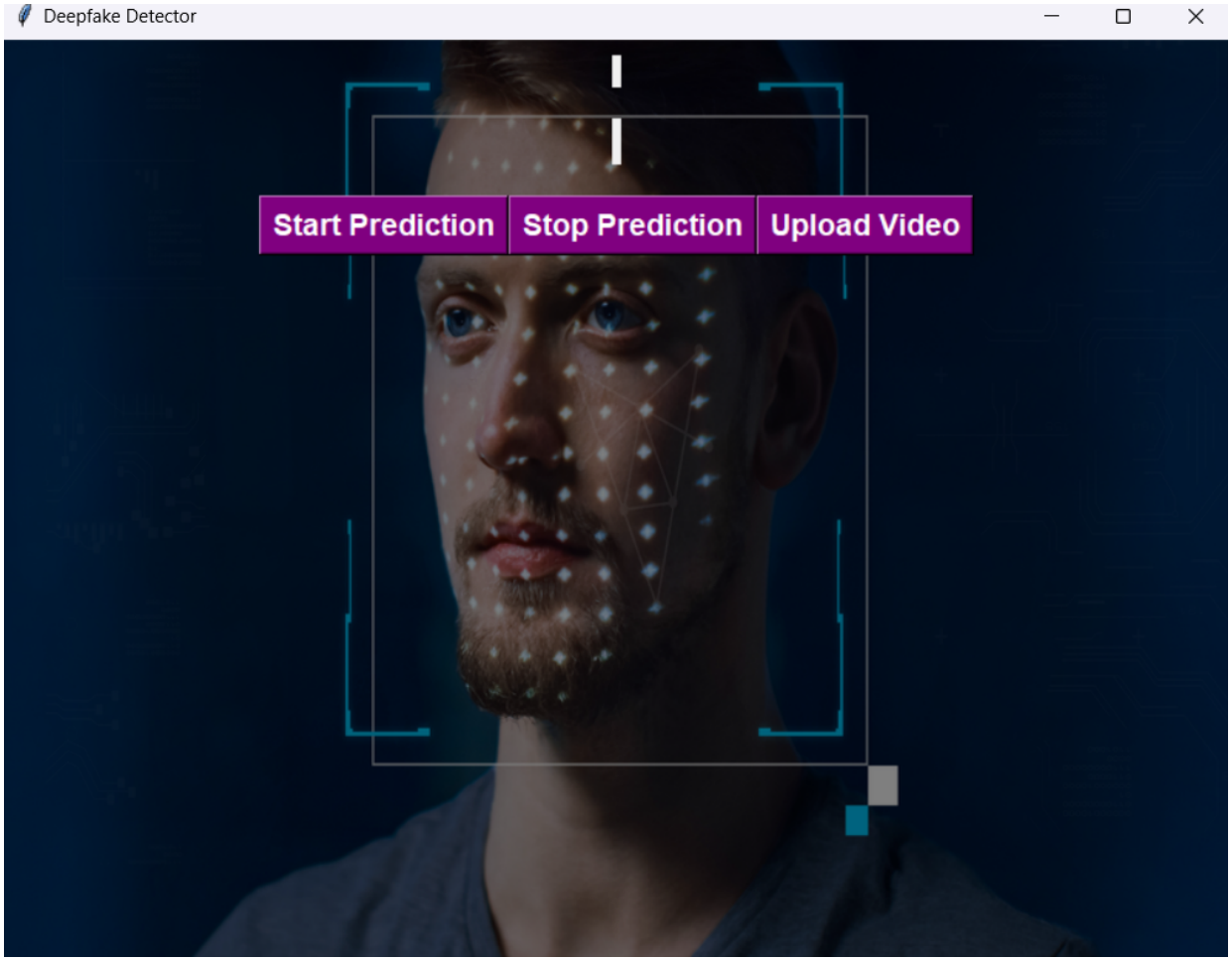

Figure 8: UI

IMPLEMENTATION DETAILS

Data Loading and Preprocessing

- Cropping the face part, trimmed to the first 60 frames.
- Custom Dataset Class VideoDataset
  - Extracting a fixed number of frames (sequence_length) from each video.
  - Applying transformations to each frame for preprocessing (resizing, normalization).
  - Labeling frames as fake (0) or real (1) based on metadata.
- Data Loader: to batch and shuffle the dataset for training and validation.

Transformations

- Convert the input to a PIL Image
- Convert the PIL Image to a PyTorch tensor.
- Resize the tensor
- Normalize the tensor by subtracting mean values and dividing by standard deviation values.

IMPLEMENTATION DETAILS

Model Architecture

- CNN + LSTM Approach: Utilizes a ResNext50 CNN model for feature extraction from frames, followed by an LSTM network to analyze temporal dependencies between frames.
- Dropout and Activation: Incorporates dropout for regularization and LeakyReLU activation for non-linear transformations.

Hyperparameters

- Sequence Length 20
- Epoch 50
- Learning rate 0.001
- Dropout 0.4

RESULTS

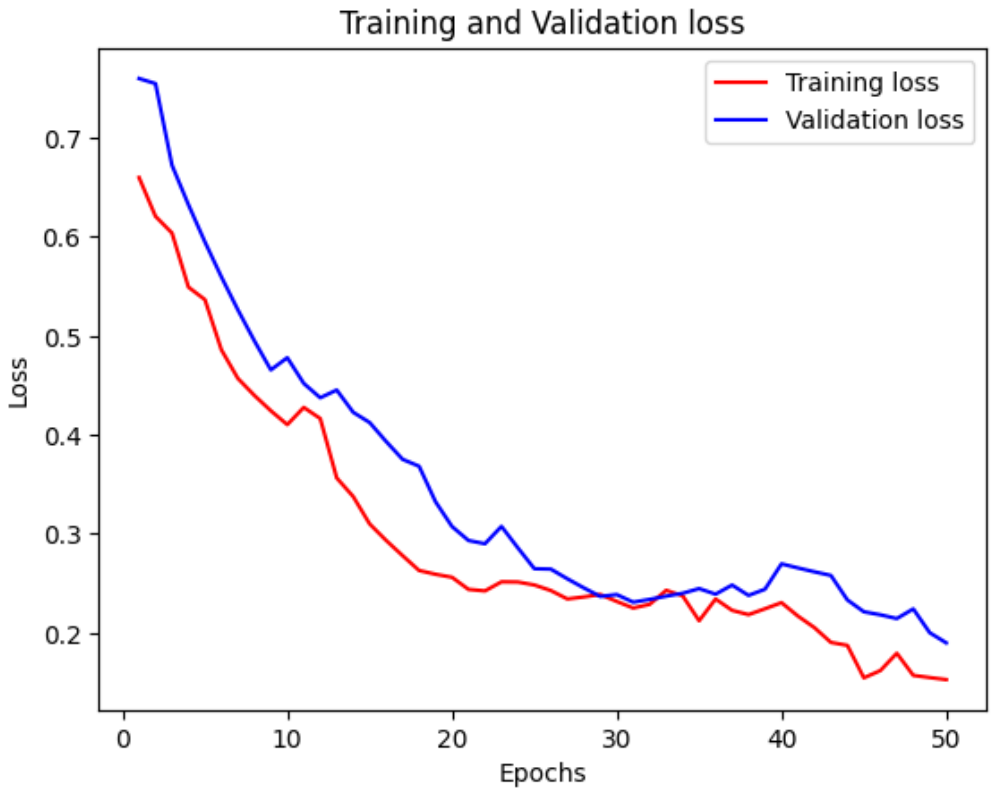

Figure 9: Loss

RESULTS

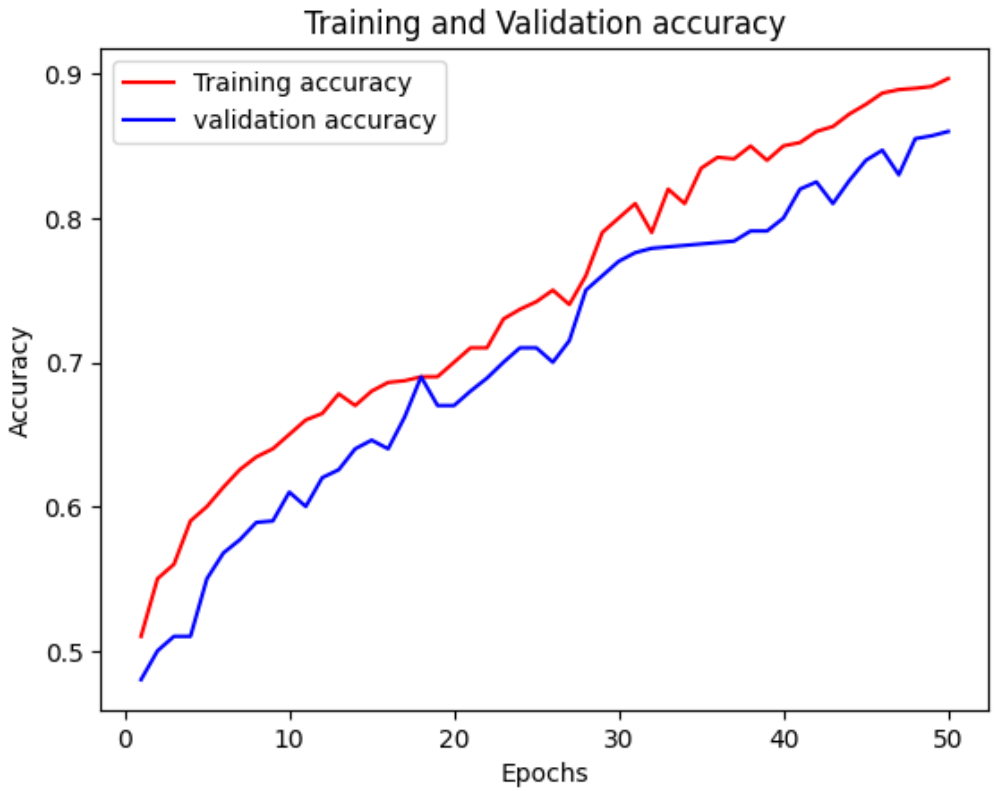

Figure 10: Accuracy

RESULTS

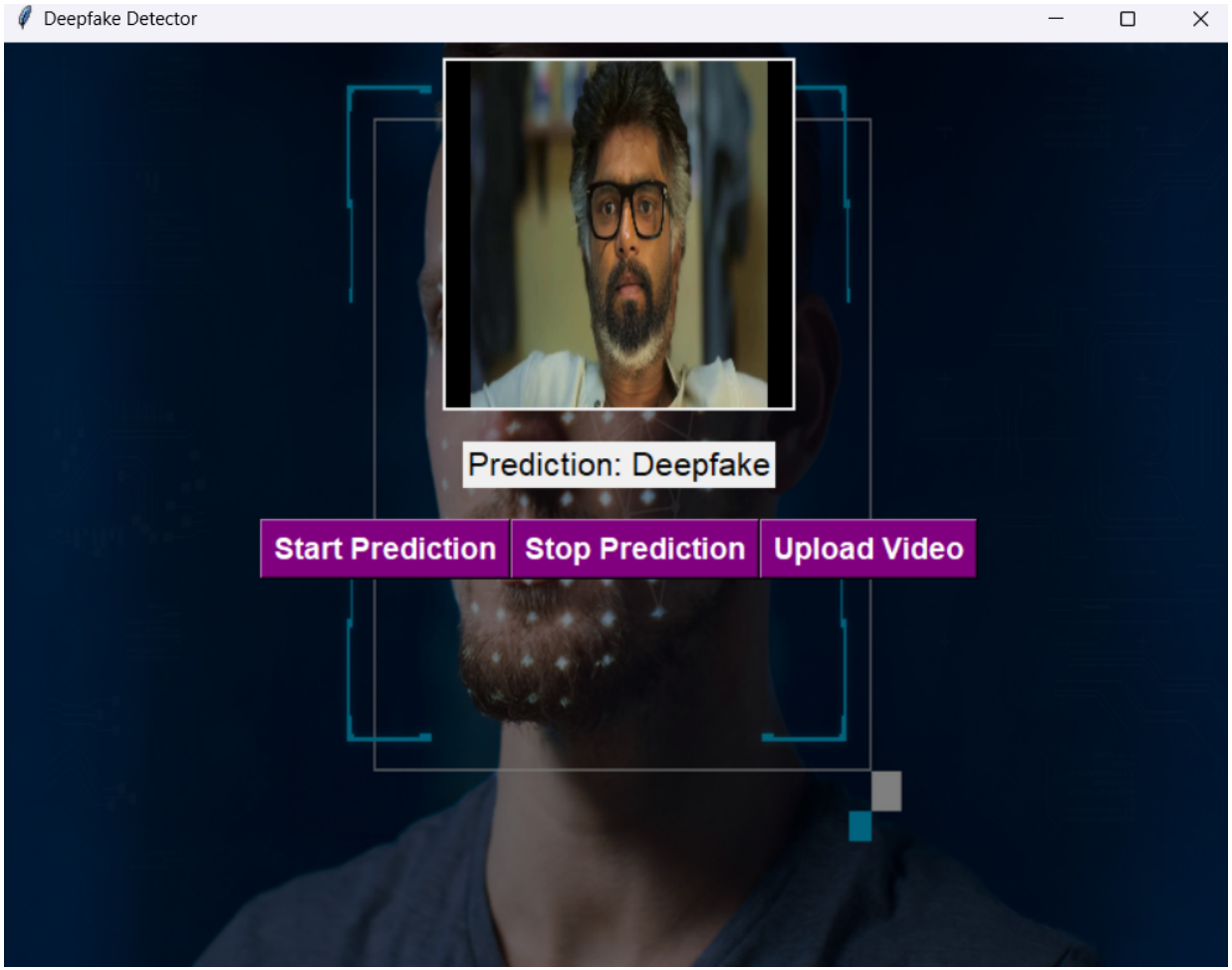

Figure 11: UI Prediciton

RESULTS

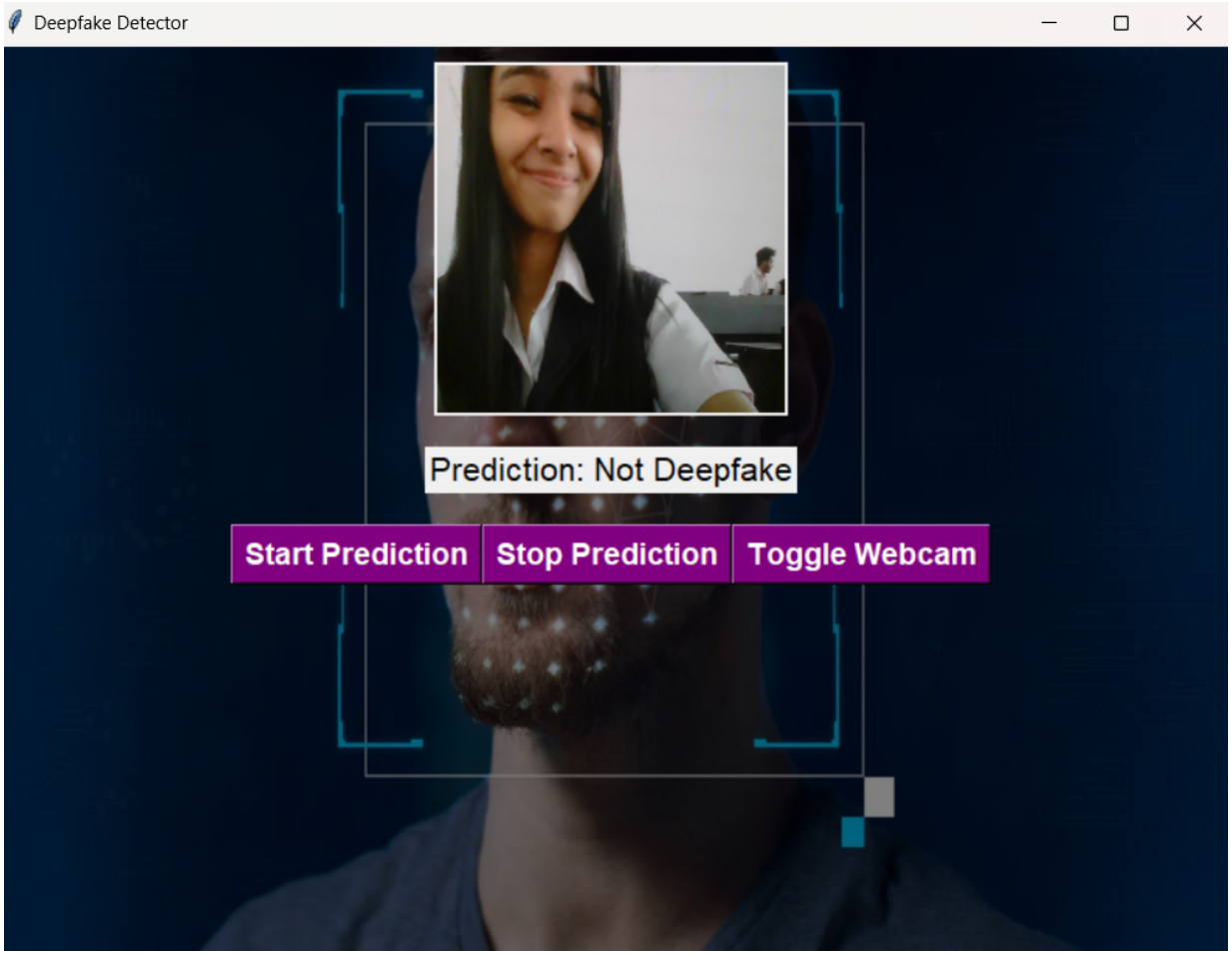

Figure 12: UI Prediciton

CONCLUSION

- Achieving 86% testing and 89% training accuracy.
- Demonstrating minimal errors with 0.19 testing and 0.15 training loss.
- Robust performance with small gaps between training and testing metrics, indicating good generalization.
- Successfully meeting project objectives with effective classification capabilities.
- Suggests promising potential for real-world deployment, with opportunities for further optimization.

References I

[1] M. Groh, Z. Epstein, C. Firestone, and R. Picard, "Deepfake detection by human crowds, machines, and machine-informed crowds," *Proceedings of the National Academy of Sciences*, vol. 119, no. 1, p. e2110013119, 2022.

[2] D. Siegel, C. Kraetzer, S. Seidlitz, and J. Dittmann, "Media forensics considerations on deepfake detection with hand-crafted features," *Journal of Imaging*, vol. 7, no. 7, p. 108, 2021.

[3] A. M. Almars, "Deepfakes detection techniques using deep learning: a survey," *Journal of Computer and Communications*, vol. 9, no. 05, pp. 20–35, 2021.

[4] M. Koopman, A. M. Rodriguez, and Z. Geradts, "Detection of deepfake video manipulation," in *The 20th Irish machine vision and image processing conference (IMVIP)*, pp. 133–136, 2018.

References II

[5] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious deepfakes: Survey, battleground, and horizon," *International journal of computer vision*, vol. 130, no. 7, pp. 1678–1734, 2022.

[6] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts. arxiv 2018," *arXiv preprint arXiv:1811.00656*, 1811.

[7] D. A. Coccomini, G. K. Zilos, G. Amato, R. Caldelli, F. Falchi, S. Papadopoulos, and C. Gennaro, "Mintime: Multi-identity size-invariant video deepfake detection," *arXiv preprint arXiv:2211.10996*, 2022.

[8] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv preprint arXiv:2102.11126*, 2021.

[9] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," in *International conference on image analysis and processing*, pp. 219–229, Springer, 2022.

References III

[10] V. R. B. M. B. S. Dr. CH.V. Phani Krishna, Sowmya Arukala, "Deepfake detection using lstm and resnext," *Journal of Engineering Sciences*, pp. 1–9, 2022.

[11] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of cnns," in *2020 25th international conference on pattern recognition (ICPR)*, pp. 5012–5019, IEEE, 2021.

[12] S. Tariq, S. Lee, and S. S. Woo, "A convolutional lstm based residual network for deepfake video detection," *arXiv preprint arXiv:2009.07480*, 2020.

[13] R. Rafique, R. Gantassi, R. Amin, J. Frnda, A. Mustapha, and A. H. Alshehri, "Deep fake detection and classification using error-level analysis and deep learning," *Scientific Reports*, vol. 13, no. 1, p. 7422, 2023.

References IV

[14] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–7, IEEE, 2018.

[15] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "Id-reveal: Identity-aware deepfake video detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15108–15117, 2021.

[16] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," *arXiv preprint arXiv:1611.05431*, 2017.

THANK YOU